Data Warehousing 2017-18
Project Assignment

v0

November 28, 2017

WARNING: THIS MIGHT NOT BE THE CURRENT VERSION OF THE ASSIGNMENT, CHECK FOR THE LAST VERSION ON THE AULAWEB MODULE OF THE CLASS

Instructions for installing docker and running the container are provided for Hive and Spark SQL proposed Lab activities (i.e., exercise 0).

General rules for project development and subsmission:

- The project must be developed by all the students taking the Data Warehouse course for 9 or 12 ECTS, or integrating previous DW editions passed in previous academic years with additional 6 ECTS.

- The project must be developed by teams of **two or three** students. Individual projects will be accepted in some exceptional cases, please contact the instructor in advance if you plan to work on the project alone. The intention to develop the project as a team must be communicated via email to the instructor when starting working on the project.

- The project must be completed and uploaded on Aulaweb at least 48 hours before the oral exam/project discussion. Students developing the project in teams can sustain the oral exam/discuss the project on different dates.

What should a submission include? Each student/team must upload on Aulaweb a zip file containing all the developed code and a pdf file with the required documentation for each of the requests below.

The code must be adequately structured in directories and instructions on how to use the code included in the zip file must also be included.

The documentation must include all the assumptions, conceptual and logical schema adequately commented and motivated, description and motivations of the approaches followed in the various steps, a critical analysis (and comparison when appropriate) on the use of tools/frameworks and relevant outcomes.

Please also include in the documentation a rough estimate of the effort (in terms, e.g., of number of hours) devoted to the project overall, and on the various parts below.

Which operational data sources should be used? Two different data sources are proposed, you may decide to work *on one of them*.

- a ”Rise Against Hunger” dataset (borrowed from the Teradata University Network Student Data Challenge 2017 `www.teradatauniversitynetwork.com`)
- a ”DVD rental” dataset (borrowed from PostgreSQL sample databases `www.postgresql.org`).

Since the DVD rental dataset is much simpler, projects based on this dataset will be evaluated with maximum grade ***24***. (The maximum grade for projects with the Rise Against Hunger dataset is ***32***).

Projects on different data sources may be possible. In case you have a proposal, directly contact the instructors to check its suitability (before starting working on it).

1 Operational data sources inspection and profiling

For each dataset, you will find in the corresponding folder the dataset itself and some documentation on the schema and/or on the interesting business questions. The operational data sources consist in OLTP data.
For data inspection and profiling, besides PostgreSQL queries, you may rely on Trifacta www.trifacta.com, just uploading the csv files for the Rise Against Hunger dataset.[1]
Include in the documentation any insight you get from source inspection and profiling.

2 Data warehouse conceptual design

Produce a data warehouse conceptual design according to the DFM notation. Identify the fact(s) of greatest interest for analysis, starting from the analysis of operational sources and making some assumptions (that must be made explicit) about the workload. You may rely on the Indyco www.indyco.com tool for drawing the fact schema(s). Motivate your design decisions and discuss dinamicity in dimensions.

3 Data warehouse ROLAP logical design

Starting from the conceptual design above and possibly refining the assumptions about data volumes and workload, develop a ROLAP logical design for the data warehouse. The design must include secondary events (i.e., views). Define your fact and dimension tables in PostgreSQL www.postgresql.org and populate them with data from the operational sources. Any ETL approach is allowed, just motivate and illustrate in the documentation your choices, and include all the relevant files in the zip.

[1] Trifacta free version does not allow connections with a database, so if working on the DVD rental dataset you can either export the content of some tables in csv or resorting to other profiling tools if you want to connect to the database.

4 OLAP Queries

Specify in PostgreSQL the queries corresponding to the workload. Moreover, referring to the specific OLAP extensions of PostgreSQL for windows and window functions, specify at least a query for each category below

- Comparison of detailed and summarized data [window partitioning]
- Computing rankings [window ordering]
- Computing cumulative totals [window framing]
- Computing mobile aggregates [window framing]

5 Hive

Import your data warehouse (at least a relevant portion of it) in Hive and run the OLAP queries in the workload in Hive `hive.apache.org`. Specify at least five new relevant queries and run such queries and the ones in the workload on Hive.

6 SparkSQL

Define the Spark SQL `spark.apache.org/sql/` DataFrames corresponding to your data warehouse (at least a relevant portion of it) and create them starting from a connection to the PostgreSQL server. Specify at least five new relevant queries as operations on data frames and run them. Now submit the corresponding SQL queries to your data frames. Run the SQL queries above and the queries in the workload on the data frames.

7 Tableau

Identify the five most relevant outcomes of your analysis and provide suitable reports in Tableau `www.tableau.com` for communicating this outcomes. Include in the documentation the graphics and discuss the identified outcomes.